dual lab

PDF

PDF Days Online 2021

WCAG or PDF/UA: what's the difference

Boris Doubrov, Dual Lab

Overview

PDF Days Online 2021

- Both WCAG and PDF/UA seem to have the common goal. But they are structured in a very different manner
- PDF/UA is a very typical PDF substandard, not much different from PDF/A, PDF/X or PDF/VT in its approach to specify file format requirements
- On the contrary, WCAG is way more general and has vaguely defined projection to the file format requirements for PDF, as it
  - includes many content and processor requirements
  - implicitly assumes client / server infrastructure
- The aim of this talk is to analyze the differences and try to set up some bridge between these two standards
- This all leads to very natural, but so far open question: what it means for a PDF document to be WCAG 2.x (Level A, AA, AAA) compliant

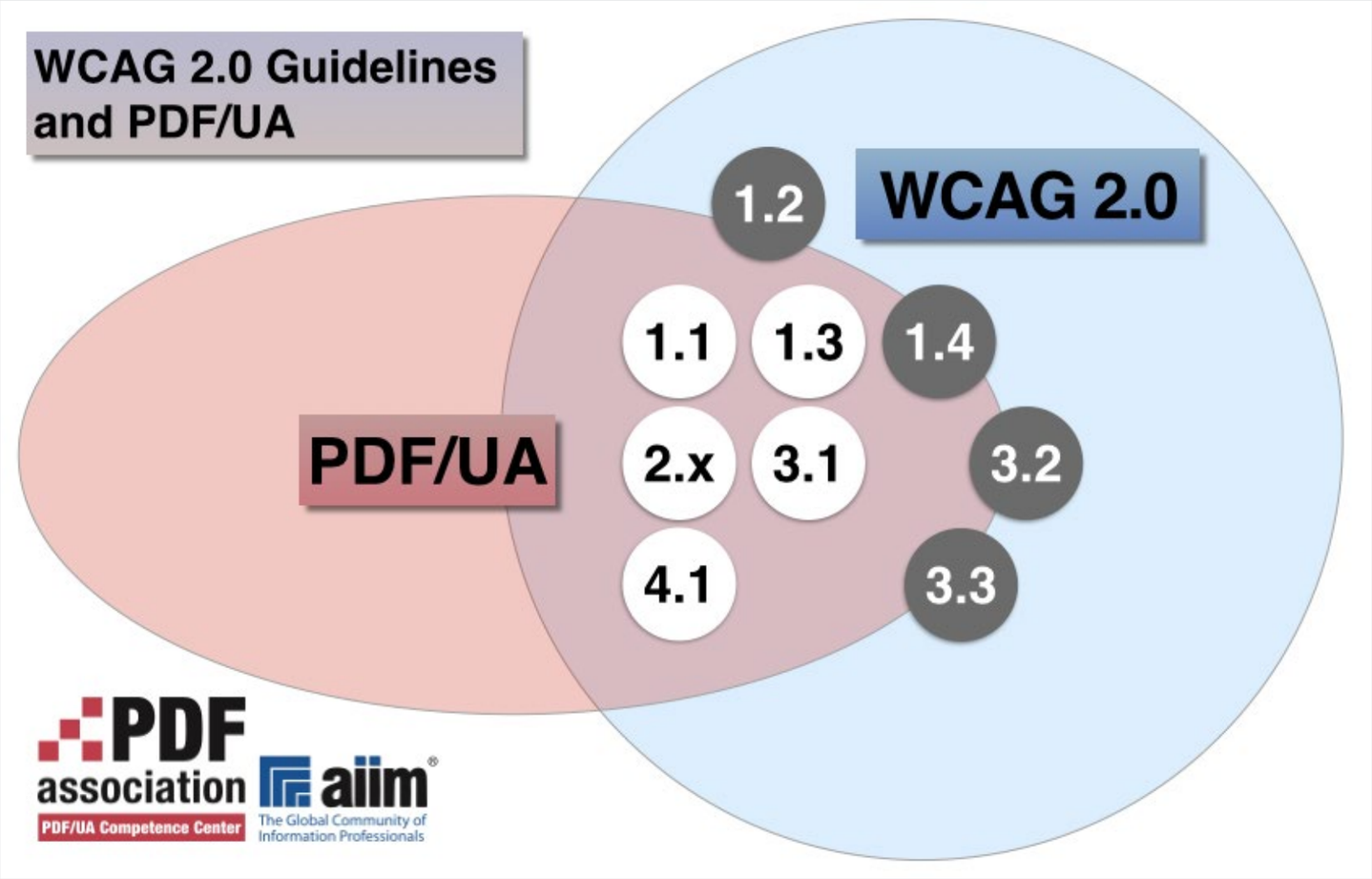
PDF/UA

- ISO 14289-1:2014 “Electronic document file format enhancement for accessibility — Part 1: Use of ISO 32000-1 (PDF/UA-1)”
- ISO 14289-2:202x based on ISO 32000-2 (PDF/UA-2) is in working drafts right now.
- Without purchasing an ISO document one can understand the requirements of PDF/UA-1 by looking at the so-called Matterhorn Protocol published by PDF Association: https://www.pdfa.org/resource/the-matterhorn-protocol/
- Lists all PDF/UA-1 checkpoints separating them into Machine and Human verifiable. The latter require (subjective) human opinion

WCAG

- Web Content Accessibility Guidelines (WCAG) is an initiative of W3C providing the standard for web content accessibility
- Current version WCAG 2.1: https://www.w3.org/WAI/standards-guidelines/wcag/ . Version 2.2 is scheduled to be published in 2021.
- Initially aimed at Open Web (HTML, CSS, JavaScript)
- Formulated via the hierarchy of Principles, Guidelines, Success criteria and accompanied with various techniques ranging from general to technology specific
- Includes PDF techniques, but only on a marginal level comparing to Web

WCAG versus PDF/UA at a glance



https://www.pdfa.org/infographics-pdfua-and-wcag-2-0/

Problems with WCAG's PDF techniques

PDF Days Online 2021

- First, there are very few of them: 60 HTML techniques, 32 CSS techniques, 19 WAI-ARIA techniques and only 23 PDF techniques
- PDF techniques for WCAG are almost 10 (!) years old. The examples are based on MS Word 2007 and Acrobat 9 !!!
- None of the 100+ failure criterions and none of 200+ general techniques mention PDF
- They don't reference either PDF/UA or Matterhorn Protocol
- Sometimes they are simply misleading and don't match similar techniques for HTML
- See https://www.w3.org/TR/WCAG20-TECHS/pdf

Examples of misleading PDF techniques

- PDF5 “Indicating required form controls in PDF forms” mentions custom validation scripts
  - However, there is no interface between PDF JavaScript and the accessibility API. Thus, there is no way to report form validation errors in an accessible manner
- PDF10 “Providing labels for interactive form controls in PDF documents” describes how to provide labels for input fields
  - Unfortunately, these are alternative descriptions rather than labels.
  - For example, WAI-ARIA clearly differentiates between **aria-labelledby** and **aria-describedby** properties.
- PDF13 “Providing replacement text using the /Alt entry for links in PDF documents” describes how to specify Alt entry for links
  - but does not discuss the case when the description can be automatically derived from the text under the link.

Problems matching WCAG with PDF/UA

PDF Days Online 2021

- Sometimes difficult to map specific WCAG Success Criteria to PDF/UA-1
- Vice versa, PDF/UA-1 inherits some low-level font requirements from other PDF standards such as PDF/A and PDF/X, which have very little relevance for accessibility and are absent in WCAG
- Accessible interactivity is plays one of the central roles in WCAG, but is underspecified in PDF/UA-1
- PDF/UA-1 leaves out important content requirements such as, for example, contrast ratio text, which are a part of WCAG 2.1 Success Criterion 1.4.3: Contrast (Minimum)
- PDF/UA-1 says very little about bookmarks, page labels and other navigational features of PDF, which are outside of the imaging model

Ongoing standardization activities

PDF Days Online 2021

- PDF 2.0 already published and formalizes structure tree schema in PDF 2.0
- PDF/UA-2 specification (based on PDF 2.0) is under development
- New specification (ISO 32005) to combine both PDF 1.7 and PDF 2.0 tag sets in a single PDF
- New work in progress on defining the notion of “well tagged” PDF
- Multiple activities of PDF Association around accessibility

PDF Association leadership

PDF Days Online 2021

- PDF/UA Technical Working Group (TWG) – all questions around the exiting (PDF/UA-1) and future (PDF/UA-2) PDF/UA standards
- PDF Reuse TWG – development of the “well tagged” PDF requirements
- Deriving PDF from HTML TWG – algorithm for deriving HTML from PDF in a predictable manner
- PDF Accessibility Liaison Working Group (LWG) – development accessible practices and potentially new PDF techniques for WCAG
- LaTeX Project LWG – accessible PDF output from LaTeX sources targeting scientific publications
- More details at: https://www.pdfa.org/industry-drives-tagged-pdf-forward

Free and open-source validators

PDF Days Online 2021

- PAC 2021 (free, but not open source) includes WCAG checks along with PDF/UA-1 ones: https://pdfua.foundation/
- CommonLook's PDF validator (free, but does require Acrobat Pro): https://commonlook.com/accessibility-software/pdf-validator/
- http://checkers.eiii.eu/en/pdfcheck/ and https://pave-pdf.org/ go back to a EU-funded project EIII completed in 2014 and never updated since that time
- https://pdfchecker.nl (led by Logius) based on veraPDF engine and a validation model that includes heuristics for checking human rules of Matterhorn Protocol

pdfchecker.nl approach

- Some PDF/UA-1 errors are not reported to the user:
  - Missing PDF/UA-1 identification in the document metadata
  - Missing CIDSet / CharSet entries in font descriptors
- Some of the rules are marked as critical. They normally result in a cascade of other errors:
  - Missing Tagging
  - Missing Language identification of the document
  - Missing or broken Unicode mapping for text within real content

pdfchecker.nl approach

- Many heuristic / human checks are included on the appropriateness of the logical structure:
  - Example: checks that tables, lists, captions, headings are marked accordingly
  - These checks are naturally subjective, come with some probability estimates and provided the end user as warnings for further inspection
- Overall associability score is computed
- Human-friendly explanations are provided (NL, EN)
- True open-source with all code at https://github.com/verapdf

Future of WCAG compliance for PDF

PDF Days Online 2021

- Formalizing what it means to say that PDF document is WCAG compliant, at least on the level of Matterhorn Protocol and at least on a basic level
- Building consistency between different validators via test corpora and canonical samples
- More effort tm map WCAG’s content requirements to PDF in a more explicit manner
- Accepting that even a partial compliance to WCAG is extremely important and may be a game changer